US006493637B1

(12) **United States Patent**
Steeg

(10) **Patent No.: US 6,493,637 B1**
(45) **Date of Patent: Dec. 10, 2002**

(54) **COINCIDENCE DETECTION METHOD, PRODUCTS AND APPARATUS**

(75) Inventor: **Evan W. Steeg**, Kingston (CA)

(73) Assignee: **Queen's University at Kingston**, Kingston (CA)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/404,714**

(22) Filed: **Sep. 24, 1999**

Related U.S. Application Data

(63) Continuation of application No. PCT/CA98/00273, filed on Mar. 23, 1998, now abandoned.
(60) Provisional application No. 60/041,472, filed on Mar. 24, 1997.

102(e)

(51) **Int. Cl.**^{7} **G06F 17/30**
(52) **U.S. Cl.** **702/19**
(58) **Field of Search** 702/19

(56) **References Cited**

FOREIGN PATENT DOCUMENTS

GB 2 283 840 A 5/1995

OTHER PUBLICATIONS

Conner et al. Proc. Natl. Acad. Sci. USA, 80, 278–282, Jan. 1983.*

Steeg et al., "Introduction of Specific Point Mutations into RNA Polymerase II by Gene Targeting in Mouse Embryonic Stem Cells: Evidence for a DNA Mismatch Repair Mechanism," *Proc. Natl. Acad. Sci. USA*, vol. 87, No. 12, National Academy of Sciences USA, Jun. 15, 1990, pp. 4680–4684.

R. Agrawal et al., "Fast Discovery of Association Rules," Chapter 12, *Advances in Knowledge Discovery and Data Mining*, pub. American Association for Artificial Intelligence, Menlo Park, California, ©1996, pp. 307–328.

D. Altschuh et al., "Correlation of Co–ordinated Amino Acid Substitutions with Function in Viruses Related to Tobacco Mosaic Virus," *J. Mol. Biol.*, vol. 193, 1987, pp. 693–707.

R. Bahadur, "A Representation of the Joint Distribution of Responses to n Dichotomous Items," Chapter 9, *Studies in Item Analysis*, ed. H. Solomon, Stanford University Press, 1962, pp. 158–175.

Carr et al., "Templates for Looking at Gene Expression Clustering," Statistical Computing & Statistical Graphics Newsletter, pp. 20–29 (Apr. 1997).

A.F.W. Coulson et al., "Protein and nucleic acid sequence database searching: a suitable case for parallel processing," *The Computer Journal*, vol. 30, No. 5, Oct. 1987, Cambridge, Great Britain, pp. 420–424.

U. Goebel et al., "Correlated Mutations and Residue Contacts in Proteins," *Proteins*, vol. 18, 1994, pp. 309–317.

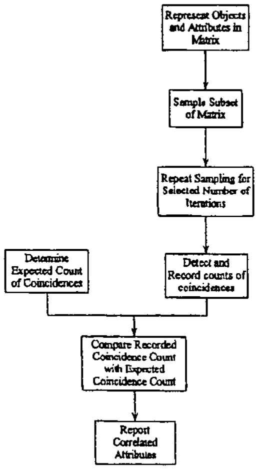
(List continued on next page.)

Primary Examiner—Michael Borin
(74) *Attorney, Agent, or Firm*—Robert H. Wilkes; Carol Miernicki Steeg; Sterne, Kessler, Goldstein & Fox PLLC

(57) **ABSTRACT**

A method and system for detecting coincidences in a data set of objects, where each object has a number of attributes. Iteratively, equally-sized subsets of the data set of sampled, and coincidences (co-occurrences of a plurality of attribute values in one or more objects in the subset) are recorded. For each coincidence of interest, the expected coincidence count is determined and compared with the observed coincidence count; this comparison is used to determine a measure of correlation for the plurality of attributes for the coincidence. The resulting set of k-tuples of correlated attributes is reported, a k-tuple of correlated attributes being a plurality of attributes for which the measure of correlation is above a predetermined threshold. The method and system (implemented on an array of processing nodes) is suitable for protein structure analysis, e.g. in HIV research.

39 Claims, 18 Drawing Sheets


Represent Objects and Attributes in Matrix
Sample Subset of Matrix
Repeat Sampling for Selected Number of Iterations
Determine Expected Count of Coincidences
Detect and Record counts of coincidences
Compare Recorded Coincidence Count with Expected Coincidence Count
Report Correlated Attributes

OTHER PUBLICATIONS

R. Guig6et al., "Inferring Correlation between Database Queries: Analysis of Protein Sequence Patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 15, No. 10, Oct. 1993, New York, New York, USA, pp. 1030–1041.

D. Heckerman, "Bayesian Networks for Knowledge Discovery," *Advances in Knowledge Discovery and Data Mining,* Chapter 11, pub. American Association for Artificial Intelligence, Menlo Park, California, ©1996, pp. 273–306.

W. Hoeffding, "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association,* vol. 58, No. 301, Mar. 1963, pp. 13–30.

Klingler, T. et al., "Discovering Structural Correlations in α–Helices," *Protein Science,* vol. 3, 1994, pp. 1847–1857.

B. Korber et al., "Covariation of Mutations in the V3 Loop of Human Immunodeficiency Virus Type 1 Envelope Protein: an Information Theoretic Analysis," *Proc. Natl. Acad. Sci. USA,* vol. 90, Aug. 1993, pp. 7176–7180.

A. Krogh et al., "Hidden Markov Models in Computational Biology: Applications to Protein Modeling," *J. Mol. Biol.,* vol. 235, 1994, pp. 1501–1531.

A.S. Lapedes et al., "Use of Adaptive Networks to Define Highly Predictable Protein Secondary–Structure Classes," *Machine Learning,* vol. 21, No. 1 / 2, Oct.–Nov. 1995, Boston, Massachusetts, USA, pp. 103–124.

C. de Marcken, "Unsupervised Language Acquisition," Ph.D. Thesis, M.I.T. (Sep. 1996) (title page, abstract, and pp. 82–93).

P. Michaud, "Clustering Techniques," *Future Generation Computer Systems,* vol. 13, Nov. 1997, NL, pp. 135–147.

R. Paturi et al., "The Light Bulb Problem," *Information and Computation,* vol. 117, No. 2, Mar. 1995, pp. 187–192.

Steeg, E. et al., "The Efficient Determination of Higher–Order Features in Protein Sequence Data" (Working Paper), Aug. 27, 1993, pp. 1–18. (FIGs. lost).

Steeg, E. et al., "Application of a Noval and Fast Information–Theoretic Method to the Discovery of Higher–Order Correlations in Protein Databases" in Proceedings of the 1998 Pacific Symposium on Biocomuting. Ed. Altman et al., World Scienctific Publishing Co., New Jersey, pp. 573–584 (Jan. 4–9, 1998).

J. Williams et al., "A Process for Detecting Correlations between Dichotomous Variables," B. Kleinmutz, ed., *Clinical Information Processing by Computer,* Holt, Rinehart and Winston, New York, 1969, pp. 100–128.

* cited by examiner

US-PAT-NO: 6493637
DOCUMENT-IDENTIFIER: US 6493637 B1
TITLE: Coincidence detection method, products and apparatus

Abstract Text - ABTX (1):

A method and system for detecting coincidences in a data set of objects, where each object has a number of attributes. Iteratively, equally-sized subsets of the data set of sampled, and coincidences (co-occurrences of a plurality of attribute values in one or more objects in the subset) are recorded. For each coincidence of interest, the expected coincidence count is determined and compared with the observed coincidence count; this comparison is used to determine a measure of correlation for the plurality of attributes for the coincidence. The resulting set of k-tuples of correlated attributes is reported, a k-tuple of correlated attributes being a plurality of attributes for which the measure of correlation is above a predetermined threshold. The method and system (implemented on an array of processing nodes) is suitable for protein structure analysis, e.g. in HIV research. 4 7 8

Brief Summary Text - BSTX (5):

The discovery of correlations among pairs of k-tuples of variables has applications in many areas of science, medicine, industry and commerce. For example, it is of great interest to physicians and public health professionals to know which lifestyle, dietary, and environmental factors correlate with each other and with particular diseases in a database of patient histories. It is potentially profitable for a trader in stocks or commodities to discover a set of financial instruments whose prices covary over time. Sales staff in a supermarket chain or mail-order distributor would be interested in knowing that consumers who buy product A also tend to buy products B and C, and this can be discovered in a database of sales records. Computational molecular biologists and drug discovery researchers would like to infer aspects of 3D molecular structure from correlations between distant sequence elements in aligned sets of RNA or protein sequences.

Brief Summary Text - BSTX (7):

Mathematical methods for determining a measure of the type, degree, and statistical significance of correlation between any two, or even three or four, particular variables are widespread and well-understood. These methods include linear and nonlinear regression for continuous variables and contingency table analysis techniques for discrete variables. However, great difficulties arise when one tries to estimate correlation--or just estimate joint or conditional probabilities--over much larger sets of variables. This intractability has one main cause--there are too many joint attribute~~-value probability density terms--and this manifests it~~self in two serious problems: (1)
7 computing and storing frequency counts over all terms, over the database, requires too much computation and
8 memory; (2) there is usually an insufficient number of database records to support reliable probability estimates
9 based on those frequency counts.

Brief Summary Text - BSTX (10):

One natural way to think about this complexity is in terms of the power set of the set of column variables. This power set forms a mathematical lattice under the operation .OR right., a "tower" corresponding to a graph whose nodes are subsets of this set of column variables. (Note that if a set has N members, the power set has 2.sup.N members). From this viewpoint, two nodes representing subsets .sigma..sub.1 and .sigma..sub.2 are connected if and only if either .sigma..sub.1.OR right..sigma..sub.2 or .sigma..sub.2.OR right..sigma..sub.1. We say that .sigma..sub.2 's node is above .sigma..sub.1 's if .sigma..sub.1.OR right..sigma..sub.2. This gives a natural meaning to the term "higher-order", as appearing higher up the tower. We call the bottom, the null set node, the 0th tier; the single column terms from the first tier, and so on.

Brief Summary Text - BSTX (15):

A comprehensive probabilistic model of the database must be able to specify probability estimates for ##EQU4##

Brief Summary Text - BSTX (17):

Clearly the models can become intractably huge. What about the space of possible models through which a modelling/learning procedure must search? Consider a latent-variable model, which seeks to explain correlations between sets of observable variables by positing latent variables whose states influence the observables jointly. Since each model must specify a set of k-tuples of variables, and there are exp(2, 2.sup.N) (i.e., 2 to the power 2.sup.N) such sets, there are exp(2, 2.sup.N) possible models in the worst-case search space.

Brief Summary Text - BSTX (18):

Various methods for determining a measure of higher-order probabilities will circumvent the combinatorial explosion through severe prior restrictions on the width k (See FIG. 3a), the locality (FIG. 2a), the number, or the degrees of correlation of the higher-order features sought, and on the kinds of models entertained (See FIG. 4a).

Brief Summary Text - BSTX (21):

It is the feature detection and data mining applications that are most relevant to the present invention. However, some of the most successful ways to estimate a full higher-order joint probability distribution of a database require the specification of exactly those higher-order terms which represent high correlations among sets of k.gtoreq.2 variables and invoking maximum entropy assumptions, and therefore the current invention is aimed at those applications as well.

Brief Summary Text - BSTX (23):

Various mathematical and computational methods have been proposed and used to estimate higher-order probabilities, to detect correlations, and to model higher-order database relationships. All such prior methods either perform a global, sometimes exhaustive search through all possible k-tuples of variables, which is too costly, or they avoid the complexity altogether by limiting their search to only k-tuples of a specific fixed, small size k. (Often, k=2 so only pairwise correlations are ever considered).

Brief Summary Text - BSTX (25):

Assuming Independence between Attributes. The easiest way to avoid the complexity of higher-order correlations is just to pretend that they do not exist. Many of the algorithms and computer programs, historically dominant in some fields of application of the current method, simply construct and use a model of the data in which all variables, all attributes, are independent. For example, the modelling of DNA and protein sequences, in computational molecular biology, is often done with consensus sequences and profiles, which assume incorrectly that the different base or amino acid residue positions are independent. Reliance on such models can obscure crucial functional and structural insights into the DNA or proteins being modelled.

Brief Summary Text - BSTX (26):

Prior Limits on k. One proposal for Gibbs models of databases is based on the use of Gibbs potentials, and it proposes a hashing method for calculating these special terms. Each kth-order potential requires an estimation of a kth-order joint probability density as well as some number of lower-order (typically k-/th-order) densities. The asymptotic time complexity of Miller's pattern-collection subroutine, the major component of the potential calculation, is, when interpreted in our terminology: ##EQU5##

Brief Summary Text - BSTX (27):

where K=k.sub.max is the highest order of features for which one will search and by which one will represent database objects. This exponential blow-up prevents one from searching for higher-order features (HOFs) of any order k much higher than 4 or 5 in databases with hundreds of attributes.

Brief Summary Text - BSTX (31):

Hidden Markov Models. Hidden Markov Models (HMMs) have been used widely and with increasing success in recent years, in both automatic speech recognition and in the modelling of protein, DNA, and RNA sequences.

Brief Summary Text - BSTX (32):

Although some groups have reported significant success in modelling protein sequence families and continuous speech data with HMMs, nonetheless there are great improvements to be made in learning time and model robustness by the "hardwiring" of pre-selected higher-order features into HMMs. (This has been investigated for HMM-like recurrent neural networks, in different domains).

Brief Summary Text - BSTX (33):

Some of the same reasons why HMMs are very good at aligning the protein sequences or recorded utterances in the first place, using local sequential correlations, make such methods less useful for finding the important sequence-distant correlations in data that has already been partially or completely aligned. The phenomenon responsible for this dilemma is termed "diffusion".

Brief Summary Text - BSTX (34):

A first-order HMM, by definition, assumes independence among sequence columns, given a hidden state sequence. Multiple alternative state sequences can in principle be used to capture longer-range interactions, but the number of these grows exponentially with the number of k-tuples of correlated columns.

Brief Summary Text - BSTX (35):

The Agrawal et al. Method for Discovery of Association Rules. This method was developed in perhaps the purest data mining context, the automatic extraction of knowledge-base rules from databases. It considers a database of M transactions (objects, rows) and N items (attributes, columns) and seeks to extract rules of the form a=>b. It therefore seek pairs of attributes a, b such that "transactions that contain a tend to contain b", hence those pairs with high values for p(b.vertline.a). "People who buy CD players tend to buy CDs", is just one example suggesting the potential commercial interests in such methods. (More generally, one can search for sets of attributes with high p(b.sub.1, b.sub.2, . . . , b.sub.k.vertline.a.sub.1, a.sub.2, . . . , a.sub.j)).

Brief Summary Text - BSTX (43):

In a first aspect the present invention provides a coincidence detection method for use with a data set of objects having a number of attributes. The base method includes the following steps: representing a set of M objects in terms of a number N.sub.A of variables ("attributes"), where an attribute is said to occur in an object if the object possesses the attribute; sampling a subset of r.sub.i out of the M objects, for each iteration among a predetermined number of iterations; detecting and recording coincidences among sets of k of the attributes in each sampled subset of objects, a coincidence being the co-occurrence of 1.ltoreq.k.ltoreq.N.sub.A attributes in the same h.sub.i out of r.sub.i objects in the sampled subset, where 0.ltoreq.h.sub.i.ltoreq.r.sub.i ; determining an expected count of coincidences for any set of k attributes and a predetermined number of iterations of sampling and coincidence-counting as described above, the determining being performed before sampling and collecting, at the same time or after sampling and collecting; comparing, for any set of k attributes and number of iterations of sampling and coincidence-counting, the observed count versus the expected count of coincidences, and from this comparison determining a measure of correlation (or association, or dependence) for the set of k attributes; and reporting a set of k-tuples of correlated attributes, where a k-tuple of correlated attributes is a set of k of the N.sub.A attributes which have been determined by this process to have a value for a chosen correlation measure above a predetermined threshold value.

Brief Summary Text - BSTX (44):

In a second aspect the invention provides a coincidence detection method for use with a data set of objects having a number of attributes, the method comprising the steps of: sampling a subset of the data set for a predetermined number of iterations, each iteration the sampled subset of the data set having for each object the same subset of attributes; detecting, and recording counts of, coincidences in each sampled subset of the data set, a coincidence being the co-occurrence of a plurality of attribute values in one or more objects in a sampled coincidence over all of the sampled subsets.

Brief Summary Text - BSTX (46):

In a third aspect the invention provides a method for visual exploration of a data set of objects having a number of attributes, the method comprising the steps of: sampling a subset of the data set for a predetermined number of iterations, each iteration the sampled subset of the data set having the same number of objects although not necessarily the same objects and having for each object the same subset of attributes; detecting, and recording counts of, coincidences in each sampled subset of the data set, a coincidence being the co-occurrence of a plurality of attribute values in one or more objects in a sampled subset of the data set, where the plurality of attribute values is the same for each occurrence, the detecting and recording counts of coincidences in each sampled subset of the data set being performed before, at the same time or after sampling, detecting and recording counts of coincidences in other subsets; determining an expected count for each coincidence of interest, the determining being performed before, at the same time, or after sampling, detecting and recording; comparing, for each coincidence of interest, the observed count of coincidences versus the expected count of coincidences, and from this comparison determining a measure of correlation for the plurality of attributes for the coincidence; and reporting a set of k-tuples of correlated attributes to a user through a graphical interface, where a k-tuple of correlated attributes is a plurality of attributes for which the measure of correlation is above a respective pre-determined threshold.

Brief Summary Text - BSTX (47):

In a fourth aspect the invention provides a pre-processing method for use with a data modelling unit to capture and report to the data modelling unit higher order interactions of a data set of objects having a number of attributes, the method comprising the steps of: sampling a subset of the data set for a predetermined number of iterations, each iteration the sampled subset of the data set having for each object the same subset of attributes; detecting, and recording counts of, coincidences in each sampled subset of the data set, a coincidence being the co-occurrence of a plurality of attribute values in one or more objects in a sampled subset, where the plurality of attribute values is the same for each occurrence, the detecting and recording counts of coincidences in each sampled subset being performed before, at the same time or after sampling, detecting and recording counts of coincidences in other subsets; determining an expected count for each coincidence of interest, the determining being performed before, at the same time, or after sampling, detecting and recording; comparing, for each coincidence of interest, the observed count of coincidences versus the expected count of coincidences, and from this comparison determining a measure of correlation for the plurality of attributes for the coincidence; and reporting to the data modelling unit a set of k-tuples of correlated attributes, where a k-tuple of correlated attributes is a plurality of attributes for which the measure of correlation is above a respective pre-determined threshold.

Brief Summary Text - BSTX (48):

In a fifth aspect the invention provides a correlation elimination method for use with a data set of objects having a number of attributes, the method comprising the steps of: sampling a subset of the data set for a predetermined number of iterations, each iteration the sampled subset of the data set having for each object the same subset of attributes; detecting, and recording counts of, coincidences in each sampled subset of the data set, a coincidence being the co-occurrence of a plurality of attribute values in one or more objects in a sampled subset of the data set, where the plurality of attribute values is the same for each occurrence, the detecting and recording counts of coincidences in each sampled subset being performed before, at the same time or after sampling, detecting and recording counts of coincidences in other subsets; determining an expected count for each coincidence of interest, the determining being performed before, at the same time, or after sampling, detecting and recording; comparing, for each coincidence of interest, the observed count of coincidences versus the expected count of coincidences, and from this comparison determining a measure of correlation for the plurality of attributes for the coincidence; and eliminating a set of k-tuples of correlated attributes, where a k-tuple of correlated attributes is a plurality of attributes for which the measure of correlation is above a respective pre-determined threshold.

Brief Summary Text - BSTX (51):

In a sixth aspect the invention provides a coincidence detection system for use with a data set of objects, each object having a plurality of attributes, the system comprising: means for sampling a subset of the data set for a predetermined number of iterations, each iteration the sampled subset of the data set having for each object the same subset of attributes; means for detecting, and recording counts of, coincidences in each sampled subset of the data set, a coincidence being the co-occurrence of a plurality of attribute values in one or more objects in a sampled subset of the data set, where the plurality of attribute values is the same for each occurrence, the detecting and recording counts of coincidences in each sampled subset being performed before, at the same time or after sampling, detecting and recording counts of coincidences in other subsets; means for determining an expected count for each coincidence of interest, the determining being performed before, at the same time, or after sampling, detecting and recording; means for comparing, for each coincidence of interest, the observed count of coincidences versus the expected count of coincidences, and from this comparison determining a measure of correlation for the plurality of attributes for the coincidence; and means for reporting a set of k-tuples of correlated attributes, where a k-tuple of correlated attributes is a plurality of attributes for which the measure of correlation is above a respective pre-determined threshold.

Brief Summary Text - BSTX (52):

In the system of the sixth aspect, the means for sampling a subset of the data set may comprise means for
dividing the data set into subsets for sampling. The means for detecting and recording counts of coincidences
may comprise an array of processing nodes, each processing node detecting and recording a respective
subcount of coincidences, and the means for comparing, for each coincidence of interest, said observed count of
coincidences to said expected count of coincidences may comprise means for merging said subcounts to provide
said observed count. At least one of said processing nodes may comprise a respective subarray of processing
nodes that detect and record respective subsubcounts of coincidences, and said means for merging merges said
8 subsubcounts to provide said subcounts and/or said observed count. Each processing node may comprise
memory including an input buffer for storing received subsets of the data set and an output buffer for storing the
10 subcount or the subsubcount; and a memory bus that transfers data to and from the memory.

Brief Summary Text - BSTX (53):

In a seventh aspect the invention provides coincidence detection programmed media for use with a computer and with a data set of objects having a number of attributes, the programmed media comprising: a computer program stored on storage media compatible with the computer, the computer program containing instructions to direct the computer to: sample a subset of the data set for a predetermined number of iterations, each iteration the sampled subset of the data set having for each object the same subset of attributes; detect and record counts of coincidences in each sampled subset of the data set, a coincidence being the co-occurrence of a plurality of attribute values in one or more objects in a sampled subset of the data set, where the plurality of attribute values is the same for each occurrence, the detecting and recording counts of coincidences in each sampled subset being performed before, at the same time or after sampling, detecting and recording counts of coincidences in other subsets; determine an expected count for each coincidence of interest, the determining being performed before, at the same time, or after sampling, detecting and recording; compare, for each coincidence of interest, the observed count of coincidences versus the expected count of coincidences, and from this comparison determine a measure of correlation for the plurality of attributes for the coincidence; and report a set of k-tuples of correlated attributes, where a k-tuple of correlated attributes is a plurality of attributes for which the measure of correlation is above a respective pre-determined threshold.

Brief Summary Text - BSTX (54):

In an eighth aspect the invention provides a coincidence detection system for use with a data set of objects having a number of attributes, the system comprising: a computer; and a computer program on media compatible with the computer, the computer program directing the computer to: sample a subset of the data set for a predetermined number of iterations, each iteration the sampled subset having for each object the same subset of attributes, detect, and record counts of, coincidences in each sampled subset of the data set, a coincidence being the co-occurrence of a plurality of attribute values in one or more objects in a sampled subset of the data set, where the plurality of attribute values is the same for each occurrence, the detecting and recording counts of coincidences in each sampled subset being performed before, at the same time or after sampling, detecting and recording counts of coincidences in other subsets; determine an expected count for each coincidence of interest, the determining being performed before, at the same time, or after sampling, detecting and recording, compare, for each coincidence of interest, the observed count of coincidences versus the expected count of coincidences, and from this comparison determine a measure of correlation for the plurality of attributes for the coincidence, and report a set of k-tuples of correlated attributes, where a k-tuple of correlated attributes is a plurality of attributes for which the measure of correlation is above a respective pre-determined threshold.

Brief Summary Text - BSTX (56):

In a ninth aspect the invention provides a product having a set of attributes selected by: sampling a subset of a data set representing objects versus attributes for a predetermined number of iterations, each iteration the sampled subset having the same number of objects although not necessarily the same objects and having for each object the same subset of attributes, detecting, and recording counts of, coincidences in each sampled subset of the data set, a coincidence being the co-occurrence of a plurality of attribute values in one or more objects in a sampled subset of the data set, where the plurality of attribute values is the same for each occurrence, the detecting and recording counts of coincidences in each sampled subset being performed before, at the same time or after sampling, detecting and recording counts of coincidences in other subsets, determining an expected count for each coincidence of interest, the determining being performed before, at the same time, or after sampling, detecting and recording, comparing, for each coincidence of interest, the observed count of coincidences versus the expected count of coincidences, and from this comparison determining a measure of correlation for the plurality of attributes for the coincidence, and reporting a set of k-tuples of correlated attributes, where a k-tuple of correlated attributes is a plurality of attributes for which the measure of correlation is above a respective pre-determined threshold.

Brief Summary Text - BSTX (57):

In a tenth aspect the invention provides a product defined by applying a set of rules generated from: sampling a subset of a data set representing objects versus attributes for a predetermined number of iterations, each iteration the sampled subset having for each object the same subset of attributes, detecting and recording counts of coincidences in each sampled subset of the data set, a coincidence being the co-occurrence of a plurality of attribute values in one or more objects in a sampled subset of the data set, where the plurality of attribute values is the same for each occurrence, the detecting and recording counts of coincidences in each sampled subset being performed before, at the same time or after sampling, detecting and recording counts of coincidences in other subsets, determining an expected count for each coincidence of interest, the determining being performed before, at the same time, or after sampling, detecting and recording, comparing, for each coincidence of interest, the observed count of coincidences versus the expected count of coincidences, and from this comparison determining a measure of correlation for the plurality of attributes for the coincidence, and reporting a set of k-tuples of correlated attributes, where a k-tuple of correlated attributes is a plurality of attributes for which the measure of correlation is above a respective pre-determined threshold.

Brief Summary Text - BSTX (66):

In an eighteenth aspect the invention provides a product being defined by its interaction with a set of attributes selected by: sampling a subset of a data set representing objects versus attributes for a predetermined number of iterations, each iteration the sampled subset of the data set having the same number of objects although not necessarily the same objects and having for each object the same subset of attributes, detecting, and recording counts of, coincidences in each sampled subset of the data set, a coincidence being the co-occurrence of a plurality of attribute values in one or more objects in a sampled subset, where the plurality of attribute values is the same for each occurrence, the detecting and recording counts of coincidences in each sampled subset being performed before, at the same time or after sampling, detecting and recording counts of coincidences in other subsets, determining an expected count for each coincidence of interest, the determining being performed before, at the same time, or after sampling, detecting and recording, comparing, for each coincidence of interest, the observed count of coincidences versus the expected count of coincidences, and from this comparison determining a measure of correlation for the plurality of attributes for the coincidence, and reporting a set of k-tuples of correlated attributes, where a k-tuple of correlated attributes is a plurality of attributes for which the measure of correlation is above a pre-determined threshold.

Brief Summary Text - BSTX (70):

In any of the aspects the method provided may further comprise the steps of first creating a database of transitions between system states, wherein a system state is represented by a value of a state variable, over a chosen time quantum, and presenting the database, in whole or part, as a data set such that each state to state transition set corresponds to one of M objects and so that each state variable corresponds to an attribute.

Brief Summary Text - BSTX (71):

In any of its aspects the method provided may further comprise the steps of first creating a database of states and actions covering a chosen time quantum and presenting the database, in whole or part, as a data set such that each state/action/state triple corresponds to one of M objects and so that each state variable or action type corresponds to an attribute.

Brief Summary Text - BSTX (72):

In a nineteenth aspect the invention provides a coincidence detection method for use with a data set of objects having a number of attributes represented in a matrix of objects versus attributes, the method comprising the steps of: sampling a subset of the matrix for a predetermined number of iterations, each iteration the sampled subset of the matrix having for each object the same subset of attributes; detecting, and recording counts of, coincidences in each sampled subset of the matrix, a coincidence being the co-occurrence of a plurality of attribute values in one or more objects in a sampled subset of the matrix, where the plurality of attribute values is the same for each occurrence, the detecting and recording counts of coincidences in each sampled subset being performed before, at the same time or after sampling, detecting and recording counts of coincidences in other subsets; determining an expected count for each coincidence of interest, the determining being performed before, at the same time, or after sampling, detecting and recording; comparing, for each coincidence of interest, the observed count of coincidences versus the expected count of coincidences, and from this comparison determining a measure of correlation for the plurality of attributes for the coincidence; and reporting a set of k-tuples of correlated attributes, where a k-tuple of correlated attributes is a plurality of attributes for which the measure of correlation is above a respective pre-determined threshold.

Drawing Description Text - DRTX (3):

FIG. 1a is a depiction of a power set of a set with N=6 objects, arranged as a lattice under a subset operation, representing all possible K-triples of columns from the power set.

Drawing Description Text - DRTX (4):

FIG. 1b is a depiction of the relative portions of all lattice nodes shown (dark squares) or omitted (light squares) by FIG. 1a.

Drawing Description Text - DRTX (6):

FIG. 2b is a depiction of the relative portion of all lattice nodes shown or omitted in FIG. 2a with a subset of the terms highlighted.

Drawing Description Text - DRTX (9):

FIG. 4a is a depiction of a partition of the variables of the objects of the power set of FIG. 1a. A partition is one particular and important kind of componential model of a sequence family or other aligned dataset. In a componential model, a set of N.sub.Y latent y.sub.i variables is found to "generate" or "explain" a larger set of N observable variables c.sub.i. In a partition model, N.sub.Y.ltoreq.N, each c.sub.j is generated by exactly one of the y.sub.i, and typically N.sub.Y <N. The observables corresponding to one latent variable form a kind of clique, and presumably are highly correlated with each other and relatively uncorrelated with variables outside the clique. In FIG. 4a, the observables are formed into three cliques: (C.sub.1, (C.sub.2, C.sub.5, C.sub.6), and (C.sub.3, C.sub.4).

Drawing Description Text - DRTX (21):

FIG. 14 is a diagram of a node of a hardware implementation of a preferred embodiment.

Detailed Description Text - DETX (2):

As previously set out, a base method described herein employs the steps of: representing a set of M objects in terms of a number N.sub.A of variables ("attributes"), where an attribute is said to occur in an object if the object possesses the attribute; sampling a subset of r.sub.i out of the M objects, for each iteration among a predetermined number of iterations; detecting and recording coincidences among sets of k of the attributes in each sampled subset of objects, a coincidence being the co-occurrence of 1.ltoreq.k.ltoreq.N.sub.A attributes in the same h.sub.i out of r.sub.i objects in the sampled subset, where 0.ltoreq.h.sub.i.ltoreq.r.sub.i ; determining an expected count of coincidences for any set of k attributes and a predetermined number of iterations of sampling and coincidence-counting as described above, the determining being performed before sampling and collecting, at the same time or after sampling and collecting; comparing, for any set of k attributes and number of iterations of sampling and coincidence-counting, the observed count versus the expected count of coincidences, and from this comparison determining a measure of correlation (or association, or dependence) for the set of k attributes; and reporting a set of k-tuples of correlated attributes, where a k-tuple of correlated attributes is a set of k of the N.sub.A attributes which have been determined by this process to have a value for a chosen correlation measure above a predetermined threshold value.

Detailed Description Text - DETX (3):

An alternative base method can include the following steps: sampling a subset of the data set for a predetermined number of iterations, each iteration the sampled subset of the data set having for each object the same subset of attributes; detecting, and recording counts of, coincidences in each sampled subset of the data set, a coincidence being the co-occurrence of a plurality of attribute values in one or more objects in a sampled subset of the data set, where the plurality of attribute values is the same for each occurrence, the detecting and recording counts of coincidences in each sampled subset of the data set being performed before, at the same time or after sampling, detecting and recording counts of coincidences in other subsets; determining an expected count for each coincidence of interest, the determining being performed before, at the same time, or after sampling, detecting and recording; comparing, for each coincidence of interest, the observed count of coincidences versus the expected count of coincidences, and from this comparison determining a measure of correlation for the plurality of attributes for the coincidence; and reporting a set of k-tuples of correlated attributes, where a k-tuple of correlated attributes is a plurality of attributes for which the measure of correlation is above a respective pre-determined threshold.

Detailed Description Text - DETX (5):

In the preferred embodiment it is preferred for simplicity of programming and interpretation to use a matrix where the objects are rows and the attributes are columns; however, this is not strictly required and any of the embodiments can utilize a data set of objects and attributes that are not represented in the form of a matrix by sampling subsets of the data set directly. As known to persons skilled in the art, any relational database can be easily transformed into a 2-dimensional matrix format.

Detailed Description Text - DETX (6):

The embodiments described herein lend themselves particularly well to parallel processing as the steps of detecting, recording and counting coincidences for each of the r samples can be performed simultaneously across many different samples or other subsets of the data set.

Detailed Description Text - DETX (8):

More formally, assume that we are given a database of M objects O.sub.1, O.sub.2, . . . , O.sub.M each of which is characterized by particular values a.sub.ij.di-elect cons.A.sub.j for each of N discrete-valued variables v.sub.j. A particular value for a particular variable is denoted a.sub.1 @v.sub.j. One may start with continuously-valued variables and use any of several known methods to quantize them into discrete variables. We also note that, in many applications, the same alphabet A of possible values is used for all the variables. Each object might be a particular record in a database, or may be a sample from a random source.

Detailed Description Text - DETX (9):

If the initial N variables are not binary then they can be converted into a set of N.sub.A attributes. For example, in the input listing attached in Appendix "B" each amino acid position is a variable that has 20 possibilities corresponding to the 20 naturally occurring amino acids represented by a subset of letters from the alphabet. In order to turn the variables into binary attributes, each variable becomes 20 different attributes having 1 of 2 states, such as "A" or "not A", "B" or not "B", and so on. An embodiment for representing variables of this type is included in the source code listing in Appendix "A". Other techniques for representing data as attributes could be used.

Detailed Description Text - DETX (16):

For our purposes we consider correlation in terms of deviation from statistical independence. One can compare an observed number of occurrences of some event in viewing the database versus the number of expected if an underlying hypothesis of independent variables were true. That is, the problem is: Given the table of values, for all k=2 . . . N.sub.A, return a list of all k-tuples of attributes (a.sub.i1 @c.sub.i1, a.sub.i2 @c.sub.i2, . . . , a.sub.ik @c.sub.ik) such that

Detailed Description Text - DETX (17):

for some observed behaviour of (a.sub.i1 @c.sub.i1, a.sub.i2 @c.sub.i2, . . . , a.sub.ik @c.sub.ik), for some real number threshold .theta..sub.i.epsilon.[0, 1], and some Model which underlies one's estimation or hypothesis testing method.

Detailed Description Text - DETX (18):

The sampling subprocess may be random sampling, and if random it may be subject to any of a number of possible probability distributions over the objects, including a uniform distribution. Similarly, there may be constraints on the statistical independence or dependencies between each of the T samples drawn during the operation of the method, and between each of the r objects drawn within one sample.

Detailed Description Text - DETX (23):

Modellers of very large data sets are thwarted in their attempts to compute very far into a fully higher-order probabilistic model by both the computational complexity of the task and by the lack of data needed to support statistically significant estimates of most of the higher-order terms.

Detailed Description Text - DETX (24):

The preferred embodiment computes only a subset of higher-order probabilities, and extracts a limited selection of higher-order feature ("HOFs") for construction of a database model. Efficient use can be made of limited computing resources by pre-selecting sets of higher-order features using the correlation-detection methods described herein, and building the most significant (statistically and in terms of application-specific criteria) into model-based classifiers and predictors based on existing statistical, rule-based, neural network, or grammar-based methods. The pre-selected sets of HOFs can be used to create rules for such systems. For example, a data set may be analysed using the methods set out herein to determine that if a company is filing a patent application then it should file an assignment from the inventor. This rule is then used in the system to generate assignments whenever it is determined that a company is filing a patent application. Many rule-based networks could benefit from pre-processing using the methods described herein, see for example, the System and Method for Building a Computer-Based Rete Pattern Matching Network of Grady et al. described in U.S. Pat. No. 5,159,662 issued Oct. 27, 1992; the inference engine of Highland et al. described in U.S. Pat. No. 5,119,470 issued Jun. 2, 1992; and the Fast Method for a Bidirectional Inference of Masui et al. described in U.S. Pat. No. 5,179,632 issued Jan. 12, 1993.

Detailed Description Text - DETX (26):

Later below, practice of the principles described herein using the Los Alamos HIV Database is described. In particular, the principles were applied to study of the V3 loop of envelope proteins of human immunodeficiency virus (HIV). In biochemistry and molecular biology in general, covariation of particular residues of a protein likely indicates the existence of a structural motif characterizing a region of the protein that has a functional, physiological role.

Detailed Descripti n Text - DETX (40):

The method of the current invention can be viewed as a "high-pass filter" for detection of higher-order features. Such HOFs play an important role in database modelling, machine learning, and perception and pattern-recognition. In database mining and modelling contexts, a procedure for discovery of these features might serve any of several major roles, including: 1. Preprocessing of large, complex datasets: Many of the best modelling methods, including Gibbs models, Hidden Markov Models and EM, MacKay's density networks, and related factorial learning methods from the neural network community, could be helped significantly in capturing higher-order interactions without exhaustive search or combinatorial explosion of parameter space if preceded by a fast preprocessing procedure, such as one provided by implementing the principles described herein, that found plausibly correlated variables in the database. 2. Visual exploration of large complex data sets: If coupled to even a simple graphical display interface, a procedure such as ours permits a user to view quickly (with small number of r-samples) the most plausibly interesting higher-order features in high-dimensional data. 3. Pre-conditioning and redundancy elimination: Thus far, we have stressed the utility of finding inter-attribute correlations in order to use them in the building of models; but in many optimization, learning and data-fitting applications, one requires that correlations between variables be found and eliminated, through any of a number of subspace methods like principal components analysis (PCA).

Detailed Description Text - DETX (49):

In some preferred embodiments, different values of r are used for different sequential iterations of the sampling, and/or for different subsets of the dataset processed by different processing nodes in a parallel computing embodiment. In such cases, we may say that on the ith iteration of in the ith sample, the number of objects sampled is r.sub.i. Some advantages of using different sample sizes include: the ability to try, within one run-through of the method, different values of r when one is unsure which values of r are best; and the ability to pick different values of r for different processing nodes in a parallel computing embodiment, in order to make optimal use of different processor sizes/speeds and memory sizes among the different processing nodes. An advantage of using the same, single value of r throughout a run-through of the method is the slight gain in simplicity of the program code.

Detailed Description Text - DETX (51):

Within the computer memory is stored a data structure termed the cset table, which is a means for storing the identity and occurrence count for each cset that occurs in one or more iterations within the process. The identity of a cset is a list of attributes (columns) comprising the cset; the occurrence count is a number corresponding to the number of occurrences of a cset that have been observed up to a particular iteration within the process, or at the end of all the iterations. In some preferred embodiments, the cset table is implemented as a hash table stored in a computer memory.

Detailed Description Text - DETX (54):

Observed Counts of Coincidence. The coincidences are observed, and the corresponding csets stored or updated, by means of a binning method. In each iteration, the attributed are binned, that is, places into separate subsets according to their incidence vectors 2 over the r-sample 4 for the current iteration. In this described matrix-based embodiment of the invention, these vectors act like-r-bit addresses into a very sparse subset of 2' address space. (See FIGS. 5a and 5b).

Detailed Description Text - DETX (76):

After some number of sampling iterations has been performed, the comparing of actual to expected number of coincidences may be performed for some or all recorded coincident sets. This may be done for all csets at once, or for any subsets of them at different points throughout the process. These comparisons for different csets may be performed sequentially or in parallel, or in some combination thereof.

Detailed D scripti n Text - DETX (77):

After some number of sampling iterations has been performed, the reporting of sets of correlated attributes may be performed for some or all of the recorded coincident sets that have been determined, in the comparisons, to signal significant correlations between the component attributes. This may be done for all csets at once, or for any subsets of them at different points throughout the process. These comparisons for different csets may be performed sequentially or in parallel, or in some combination thereof.

Detailed Description Text - DETX (80):

FIG. 5a provides a pictorial example of the application of this embodiment to a fictional toy dataset. Three iterations of r-sampling (for r=3) on the toy dataset are depicted, top to bottom. For each iteration, the left-hand box represents the dataset, with outlined entries representing the sampled rows. The right-hand-box represents the set of bins into which the attributes collide. For example, in the first iteration, A@1, B@2, and D@4 all occur in the first and second of the three sampled rows, so they each have incidence vector 110 and collide in the bin labeled by that binary address. Bins containing only a single attribute are ignored; and "empty" bins are never created at all. All bins are cleared and removed after each iteration, but collisions are recorded in the Csets global data structure.

Detailed Description Text - DETX (82):

Steps 5 through 21 of the pseudo-code represents the steps of the base method described herein, namely: sampling a subset of the matrix for a predetermined number of iterations, each subset of attributes being the same, detecting and recording counts of coincidences of attributes in each sampled subset, a coincidence being the occurrence of a plurality of attributes in an object in a sampled subset, where the plurality of attributes is the same for each occurrence, determining an expected count for each coincidence of interest, the determining being performed before, at the same time, or after sampling, detecting and recording, comparing, for each coincidence of interest, the observed count of coincidences versus the expected count of coincidences, and from this comparison determining a measure of correlation for the plurality of attributes for the coincidence, and reporting a set of k-tuples of correlated attributes, where a k-tuple of correlated attributes is a plurality of attributes for which the measure of correlation is above a pre-determined threshold.

Detailed Description Text - DETX (83):

Appendix "B" contains actual source code written in the Perl language for running on a Sun4 computer in the Sun UNIX operating system. Sample input data for the code listing in Appendix "B" is listed in Appendix "C" for partial amino acid sequences from V3 loop of HIV envelope proteins. The corresponding output from the code of Appendix "B" for the input of Appendix "C" is shown in Appendix "D". In order to produce the output of Appendix "D", the adjunct Perl language program listed in Appendix "E" was used for clarification and presentation from the main code listing in Appendix "B". A general flow diagram for this embodiment is shown in FIG. 6, while a general block diagram is shown in FIG. 7. The resulting report was stored in a flat file as a relatively unstructured ascii database, which was later printed; it could equally well have been sent to a printer directly or sent across a network for report to other resources.

Detailed Description Text - DETX (88):

The method may be run entirely sequentially, as in the most straightforward interpretation of the pseudocode given above, or the method may be run on parallel (vector or multiprocessor) or distributed computer systems in many possible ways. A set of computations may be run in parallel, in which each computation performs the entire program steps outlined above, but with each separate computation using a different value for r, the sample size; or each separate computation could run the same program steps with same key parameter values, but start with different initial random number seeds for the random r-sampling. Alternatively, the entire program steps outlined above could be run once, but each different r-sample could be forked off into a separate process run on different processors, where in each such process would comprise the detection and optionally recording steps, with the global cset counts later joined into the global process and global data structures. Additionally, the computation of the expected counts, and the comparisons of expected with observed counts, could be performed all at once or incrementally, sequentially or in parallel. Similarly, the reporting of the estimated correlation values can be performed for some or all of the Csets, once at the end of computation or incrementally throughout, in serial or parallel.

Detailed Descripti n T xt - DETX (93):

Many possible ways exist for the representation, storage, and accessing of the Csets data structure used during the processing of the algorithm. The Csets data may be stored and accessed via a hash table, a k-d tree, patricia tree (also called a trie), and/or in other ways, known to those skilled in the art, of storing and accessing data efficiently. Whatever data structure is chosen, the structure may be stored physically in registers, in main memory, and/or on secondary or external storage media such as magnetic disks, magnetic tape, or optical storage media.

Detailed Description Text - DETX (95):

For example, very efficient special purpose electronic (LSI or VLSI) may be used to implement the matrix representation of the current invention, by the fact that the incidence vectors of attributes are simple binary vectors, by the fact that the coincidence "bins", described earlier in one view of the current invention, correspond to "addresses" to a memory space of size 2.sup.r for each r-sample, and by the ability with current technology to design, fabricate and use special-purpose hardware for implementations of random-number generation and sampling, fast-access storage of the Csets data structures, and of the mathematical functions used in the calculation of expected count estimates and hypothesis tests and correlation estimates.

Detailed Description Text - DETX (98):

Referring now to FIG. 14, an embodiment of special purpose hardware mentioned previously is intended to exploit the potential benefits of parallelizing the execution of the algorithm. A node (defined below) divides a given data set along M (the number of rows of data) and distributes these portions to its CPs (also defined below). The CPs may be either other nodes (in a recursive definition) or may be special purpose processors developed to perform step 8 in the method as described in high-level "pseudo-code" in the previous Program Method Description of a Preferred Embodiment Section. When the results have been computed by the node's CPs, the merging step (steps 9 through 14 in the above-noted "pseudo-code" description) is performed by the node. Once the merging has been done, the results are passed back to the node's parent. If the node is the root of the tree, the complete results set is sent back to the driver that controls this hardware. The system described below can be used "off-line" from a main computer's CPU; among other possibilities for commercial marketing and use of such a system is its implementation on a special "board" or "card" that a user can purchase and install on his or her personal computer or workstation. One can also envision the use of one or a number of such special subsystems on a local area network or a "supercomputer" installation. The described embodiment represents only one of many possible ways, as will be understood by those skilled in the art, to parallelize the methods described herein.

Detailed Description Text - DETX (100):

A diagram of a node is shown in FIG. 14 with compute processors (CPc). The node includes the following: A bank of memory where input to be sent to the CPs is stored (the input buffer) and where results found by the CPs will be stored (the output buffer). A memory bus divided into control, data and address buses used to arbitrate communication on the bus itself as well as being the vehicle for data transfer. A set of bit flags and a small additional portion of memory (LastOut). LastOut is the address of the section in the output buffer that was last written to. The two bit flags are used by the merge and I/O processors to determine what state they each are in. An array of size J of compute processors (CPs), each with their own local memory caches, which perform the discovery of coincidences. A merge processor (MG) which has its own cache of memory in which it writes the merged results of the CPs. An input/output processor (IO) whose main responsibility is to control use of the bus. A clock which is used to ensure that each element in the system runs synchronously with respect to every other element. Execution of each of the parts in the system can be thought of as running in lock-step.

Detailed Description Text - DETX (101):

Computer processors are defined as being either special processors that perform the R-sampling step of the algorithm (step 8 in the pseudo-code description and graphically in FIG. 5a. This allows the possibility of a tree structure of such nodes rather than limiting embodiments solely to a vector arrangement. For any particular choice of hardware for the memory bus, it may be the case that there is a maximally useful limit on the number of CPs per node. A tree structure allows a way around this limit.

Detailed Descripti n Text - DETX (104):

For each node, memory of size 2*J*Amax*Rmax*Nmax, where Amax is the maximal total number of iterations that can be done in the node. This memory is divided equally into the input and output buffers. Note that the size of the input for a single iteration is no greater than J*Rmax*Nmax and neither the locally-produced results nor the final merged results (formed by combining the partial results from the J CPs) can exceed this limit, so there is no risk of exceeding available memory.

Detailed Description Text - DETX (113):

As noted above, these are either nodes or are special purpose processors that compute one R-sampling step in the algorithmic description of the general method of the current invention. In the latter case, they may comprise: a processor which performs the coincidence detection in addition to the functions listed below 2*Nmax*Rmax sized local memory

Detailed Description Text - DETX (115):

Initially, a CP asserts 1 on its request wire, indicating that it is ready for data. When it sees only its response wire set to one on the following cycle, it expects to be sent the current values for R and N and then the data itself (otherwise, it waits for this to be the case). Based on the first two values, it can determine when the current input is exhausted. It then asserts 0 on its request wire and performs the binning and coincidence detection steps of the method. When these steps have been completed the CP asserts logical 1 again on its request wire, this time indicating its desire to send its results. When given permission to use the bus, it sends its coincidence set to IO. IO is responsible for managing the location for storage of this data. The output stream of the CP comprises a tally of the coincidences found followed by the coincidences (csets) themselves. The coincidences are of the form: hit count (no higher than Rmax) size (that is, the width of the cset, i.e., the number of component attributes) a size-long list of the attributes of the coincidence in form (value, position)

Detailed Description Text - DETX (120):

When MG sees that its request wire has been turned on, it knows to start receiving output data indexed by the counter into its local memory. Once this has been accomplished, MG can start the merging algorithm. The merge is done from the local memory directly into the merge buffer (C2 must have the current number of coincidences when this step is finished). When this step is completed, MG retrieves the current value of LastOut. If it is greater than C1, then Mg knows it can increment C1 and move directly on to the next output section. If C1 and LastOut are equal, then MG sets its request wire to zero. If C1 has reached A*J, then MG knows that all the results have been computed and merged (and thus, that all CPs and IO are idle) and that it should set its bit flag to one (indicating that it is finished) and start sending the contents of the merge buffer back to IO for transmission to this node's parent. The results are sent simply as the value of C2 followed by the list of coincidences stored in the merge buffer (the form of the coincidences is identical to that described in section 5 above).

Detailed Description Text - DETX (124):

IO can thus determine when no more data can be expected. Note that it is the responsibility of the driver to: divide data mining requests into sizes no greater than Amax ensure that the number of rows sent as input is evenly divisible by R ensure that Rmax and Nmax have not bee exceeded by the current data set merge all results sent back from the device

Detailed Description Text - DETX (126):

When all CPs are busy (or all available input has been exhausted), IO waits for a CP to assert 1 on its request wire which indicates that it is ready to send back results. Once this signal has been received from a CP, IO retrieves the results from the CP, stores them in the output section indexed by the counter, zeroes the bit associated with that CP, increments C1 and asserts 1 on the MG request wire. If there is unused data in the input buffer, IO sends the next available R*N set to the CP who just returned results (setting the bit for that CP to one). When C2 equals T and the bit vector contains no bits set to 1, then IO knows that it is finished and sets the IO bit flag to 1. At this point, IO goes back to the previously described wait state until it sees the MG bit flag also set to 1 (indicating that MG has finished its work). Once this occurs, IO calls an interrupt (if this node is the root of the tree) or just requests to send (if this node has another node for a parent), gives MG permission to write on the bus and then passes all data sent from MG to the parent.

Detailed Description Text - DETX (141):

The problem of finding all significant correlations among pairs or k-tuples of attributes in a database is ubiquitous in the computational sciences and in medical, industrial, and financial applications. The principles described herein include a probabilistic algorithm that has the interesting property of finding significant higher-order k-ary correlations, for all k such that 2.ltoreq.k.ltoreq.N in an N-attribute database, for the same computational cost of finding just significant pairwise correlations. Moreover, k need not, be fixed in advance in our procedure, in contrast with other known procedures. The procedure was deigned for the task of finding conserved structural relationships in aligned protein sequences, but may have more useful application in other domains.

Detailed Description Text - DETX (143):

There are interactions between sequence-distant amino acid residues in the protein chain, sometimes detectable as correlations between positions (columns) in a set of aligned sequences from a protein structural family, that play an important role in determining structure and function. Discovered correlations may represent an evolutionary history of compensatory mutations, and may provide useful features in models of protein structural/functional families, but are ignored or mishandled by most ML (machine learning) classification methods, in part because of the high computational complexity of searching for k-tuples of correlated positions.

Detailed Description Text - DETX (151):

First, there are distance geometry constraints. Secondary structure prediction, and the discovery of k-ary long-distance interactions, give evidence for presumed contacts, of the form contact(i,j) for the ith and jth amino acid residues in a protein. Using the kind of distance geometry theory developed by others (see for example, T. F. Havel, L. D. Kuntz, G. M. Crippen The Theory and Practice of Distance Geometry Bull. of Mathematics Biology v. 45 1983 pp. 665-720. and K. A. Dill, K. M. Feibig, H. S. Chan Cooperativity in Protein-Folding Kinetics Proc. Natl. Acad. Sci. U.S.A. v. 90 March 1993 pp. 1942-1946), one can derive a set of inferred contacts. One can also derive sets of inferred blocks, contacts that are forbidden by a given set of presumed or inferred contacts. Essentially, given a model of a polymer chain constrained to exist within a fixed volume, the assumption that two particular pieces are brought into contact implies that some other pieces are also brought into proximity and that still other pieces are moved further apart. Indeed, others have concluded that "considerable amounts of internal architecture (helices and parallel and anti-parallel sheets) are predicted to arise in compact polymers due simply to steric restrictions. This appears to account for why there is so much internal organization in globular proteins."

Detailed Description Text - DETX (152):

Second, as discussed throughout the previous sections, one can infer and exploit empirical relationships between local and global configurations. Local stretches of sequence, or selected non-local pairs of residues, can be found to occur, with some high probability, in particular global configurations. Heuristic rules, in whatever form, can be used to avoid large parts of conformation space. This inference of particular models of cooperativity in folding is a special case: knowledge of "rules" such as p(contact(i,j).vertline.contact(i+1,j-1))>p(contact(i,j)) can help significantly.

Detailed Description Text - DETX (167):

Tests on an HIV Protein Database

Detailed Description Text - DETX (168):

The Los Alamos HIV database contains, among other things, the amino acid sequences for the V3 loop region of the HIV envelope proteins. This region is known to have functional and immunological significance, and the discovery of sets of sites linked by evolutionary covariation might have important implications for understanding and preventing HIV infection and replication.

Detailed Description Text - DETX (169):

An earlier and smaller version of the same database was used by Los Alamos scientists in their analysis of pairwise mutual information between residues (columns)..

Detailed Description Text - DETX (171):

Results of Experiments on HIV Protein Database

Detailed Description Text - DETX (187):

Table C.8: The top thirty-five pairwise inter-column mutual information values for the HIV-V3 dataset, as estimated by our methodology as described in the main text.

Detailed Description Text - DETX (192):

In order to get a better sense of the possible meanings of these results, let us consider these inter-attribute correlations along with some inter-column correlations in the form of pairwise mutual information estimates performed in our own analysis and also by the Los Alamos group. Table C.8 displays the highest estimated mutual information values amongst all N--N=528 pairs of columns from our 33-column dataset. The estimates were obtained using a Bootstrap-like procedure in which 1000 sample data subsets of m=300 out of M=657 were drawn and run though the standard mutual information calculation. Reported in the table are therefore the mean values over the resampling and the associated standard error values. There is significant intersection between the set of column-pairs indicated by the top cset values in Tables C.6 and C.7 and those indicated by the top mutual information values in Table C.8. The correspondence between the two rankings is not perfect, for a few reasons (besides noise and simple sampling error). First and foremost, while the "suspiciousness" of a single joint-attribute combination certainly contributes to the mutual information within the corresponding set of columns the behaviour of the other symbols appearing within the columns obviously also can have great effect. Second, we note again the observed sensitivity coincidence detection results to the choice of r.

Detailed Description Text - DETX (193):

Table C.9 lists the highest statistically significant mutual information values as estimated by the Los Alamos group. We note the overlap between their list and ours, but we emphasize again that group's use of an earlier, smaller, and perhaps otherwise different database to which we did not have access.

Detailed Description Text - DETX (206):

It is therefore of great interest to biotechnology and pharmaceutical researchers to be able to consider a huge number of potentially useful compounds, but to avoid spending too many resources developing therapies based on compounds that may turn out not to be useful, safe, effective, and economically viable. The methods described herein can be used to enhance and accelerate the process of discovering good, effective compounds and of distinguishing the promising compounds from the unpromising or less promising compounds in a public or private collection of molecules or their computer database representations. They can be used effectively and contribute value in this application in many ways, by helping to understand and infer target structures and by finding ligands whose geometric, topological, electrostatic or other features make them likely candidates for effective interaction with the targets.

Detailed Description Text - DETX (207):

Application of the Principles Described Herein to Databases of Molecules and their Features

Detailed Description Text - DETX (208):

One way to represent a large number of molecular structures within a computer database (whether stored in main memory, on magnetic disk, tape, or other electronic or optical media) is in terms of "screens". Persons skilled in the art will recognize screens as binary attributes wherein a given screen, or attribute, represents the presence or absence of a particular substructure pattern, for example, sulfate group. If a set of compounds is represented with screens, then a particular compound, which we will denote by C, can be represented by a string of 1s and 0s wherein the 1s stand for those pre-defined substructure patterns that C contains and the 0s stand for those of the pre-defined substructure patterns that C does not contain.

Detailed Description Text - DETX (211):

Not only can small therapeutic compounds be represented in terms of screens and other attributes, but so can larger potentially therapeutic molecules such as DNA, RNA, peptides, proteins, carbohydrates and lipids. Target molecules can also be represented in this way. All that is required is a predefined (though possibly updated, changing, shrinking or growing) list of substructural patterns or other features deemed important by the researchers or users. For target structures, one might want to represent substructural patterns as well as their 1-dimensional linear structures ("sequence"), genetic linkage information, interactions with other proteins in disease pathways, literature citations, and so on. Sometimes a particular molecule might be listed as more than one object in a database, the different objects representing different conformations that the molecule can take.

Detailed Description Text - DETX (212):

Clearly, this use of screens and other attributes in representing compound databases can also be represented in terms of the M by N data matrix we have used to describe the working of the invention. The M by N data matrix is illustrated below in Table 1.

Detailed Description Text - DETX (214):

Steps involved in applying the methods described herein to the analysis of a molecular database include: 1. Obtain molecular database that supports discrete attribute representation for the 1D, 2D and/or 3D molecular structures of interest (or, obtain molecular database and use standard methods to produce such a representation); also use standard methods to transform sequence and other information about molecules of interest into attribute representations. 2. Present this database, in whole or part, to an embodiment of the current invention such that each compound in the database corresponds to one or more of the M objects (rows) in the embodiment's data matrix and so that each screen-represented substructure pattern corresponds to an attribute (column) of the data matrix. The additional attributes representing activity, assay results, known targets against which the compound has been used, source or means of production or storage of the compound, ownership or patent status of the compound, and so on, plus the substructure pattern attributes together comprise the N attributes (columns) in the data matrix. 3. Employ the base method above or one of the other embodiments described herein on the data matrix. 4. Direct the discovered correlated k-tuples of attributes to:

Detailed Description Text - DETX (215):

A graphical viewer, or A rule-generator preprocessor for rule-based system, or A report for users, researchers or managers, or a report-generation system, or Another computer program that performs some kind of further analysis of the compounds, sequences, or structures represented in the database, or Another computer program that performs some transformation or optimization on the database, or Another computer program that directs humans and/or robots in drug screening experiments or in design, refinement or production of therapeutic compounds.

Detailed Description Text - DETX (220):

Another example is that of finding correlated amino acid residues in that part of a drug discovery database corresponding to an aligned set of DNA, RNA or protein sequences, as discussed later herein. In this case, some of the correlated k-tuples of residues (positions) may correspond to evolutionarily conserved structural and functional relationships. Therefore the principles described herein can in this way be used to help predict or solve the structure and function of important biological macromolecules, including pharmaceutical targets such as receptors and enzymes.

Detailed Description Text - DETX (222):

Another rather different application of the principles described herein to drug discovery and medical science is obtained by considering the transpose of data matrix described above. Instead of compounds as objects (rows) and features of the compounds as attributes (columns), consider what is possible when the compounds correspond to columns and their features correspond to rows. See Table 2 below. Use of the current invention in this scenario produces correlated k-tuples of compounds in feature-space. These produced k-tuples can embody several kinds of valuable information. For example, if the features in the rows represent mostly substructural patterns (screens), then the produced k-tuples correspond to clusters of compounds. Such clustering of compound databases is very useful in high-throughput screening (HTS), with both biological/chemical assays (in vitro or in vitro) and computational assays. In HTS, it is useful and economical to assay only one or a few members of each cluster of compounds initially; then, only in the cases where a "hit" occurs (that is, a compound "passes" the "test" in the assay of biological or chemical activity) do other members of the corresponding cluster get sent through the assay.

Detailed Description Text - DETX (223):

Use of the method on the "transpose" of the molecular database shown earlier, in order to cluster the compounds in feature-space is shown in Table 2. It is now the columns that correspond to a set of molecules, compounds, molecular structures or sequences, while the rows correspond to features that may include substructural patterns, assay results or other aspects of the molecules. There are M' rows and N' columns, where perhaps M'=N and N'=M, for the original M and N described above. The value in table cell[i,j] is one (1) if molecule i has feature j and is zero (0) otherwise.

Detailed Description Text - DETX (228):

The output of the above steps, that is, a set of k-tuples of correlated attributes, can be interpreted as a set of cliques of correlated genes. For example, one might discover that one gene is "on" whenever another gene is "on". Or one might discover that when one gene G1 is in "low expression", another gene G2 is "off"; when G1 is in "medium expression", G2 is in "low expression"; and when G1 is in "high expression", then G2 is in "medium expression". Such a result might lend support to the hypothesis that G1 promotes the expression of G2, or that "G1 turns G2 on". Similarly, correlated k-tuples of genes or biological parameters might provide evidence that one gene represses, or "turns off" another gene or set of genes, and so on. All such information can be useful in building a model, for example a "boolean network", of a set of interacting genes. Such models are known to those in the art as providing valuable assistance in diagnosing, preventing and curing disease and in designing effective and economically valuable therapeutics.

Detailed Description Text - DETX (229):

The rows in Table 3 correspond to a set of time-samples (a.k.a., time points, time-slices), that is, times or periods of observance of the activity of a particular gene or gene product. The columns correspond to particular genes or gene products. The value in table cell[i,j] is one (1) if gene i is considered "on", that is, e.g., "active" or "expressed", during time j and is zero (0) otherwise. This representation and application is easily extended to situations in which the simple on/off status of a gene is replaced by a set of z distinct levels of expression, for example, as measured by observed quantities of a gene's main protein product. It is also easily extended to situations in which more than one biological parameter is used to represent the status of a single gene.

Detailed Description Text - DETX (238):

Application of the Principles Described Herein to the Discovery of Categories in Internet/Intranet Document Databases for Use in Document Search Engines

Detailed Description Text - DETX (240):

Application of the method to optimal or near-optimal topic set reduction can also be represented in terms of the M by N data matrix we have used to describe the working of the invention in other sections of this document. In one application-specific embodiment, the rows of the data matrix correspond to particular documents in the database, and the columns correspond to a proposed topic set that is intended to categorize them. (See Table 6).

Detailed Descripti n Text - DETX (241):

The rows in Table 6 correspond to documents in a database, while the columns correspond to proposed topics used to classify them. The value in table cell[i,j] is one (1) if document i mentions topic j and is zero (0) otherwise.

Detailed Descripti n Text - DETX (242):

Steps involved in applying current invention to a search for a near-optimal topic set with which to classify a set of documents include: 1. Obtain an initial topic set. The field of document search is well established and effective methodologies for the creation of such sets are known to those skilled in the art. 2. Create the database using this topic set and the set of documents that the topic set categorizes. Given the topic set, all one need do is examine each document to determine whether or not it mentions each topic. 3. Present this database, in whole or part, such that each document in the database corresponds to one or more of the M objects (rows) in the embodiment's data matrix and so that each proposed topic corresponds to an attribute (column) of the data matrix. 4. Employ the base method above or one of the other embodiments described herein on the data matrix. 5. Direct the discovered correlated k-tuples of attributes to: A graphical viewer or printer, or A rule-generator preprocessor for rule-based system, or A report for administrators or other users of the computer database query system, or a report-generation system, or Another computer program that performs some kind of further analysis of the data, for example, performing more in-depth statistical analysis (e.g., multiple regression) on the correlated variables, or Another computer program that performs some transformation or optimization on the database.

Detailed Description Text - DETX (248):

Table 7 illustrates the database upon which the base method or other embodiment described herein will be run, in the data matrix format for representing objects and attributes that have been defined and described elsewhere herein. Note that, because of the characteristics of the embodiment described herein, the number of pages used in the table need not be the entire set of all web pages. The embodiment, when run (or employed) on this table will find those topics that are frequently found in the same document together. This indicates that these topics are related in some fashion and, as the set of web pages supports their association, they may be of interest to the user as well.

Detailed Description Text - DETX (249):

The advantages are several. The computational expense of these embodiments scales linearly with respect to the number of columns in the database. In this application, the number of columns represents the number of topics associated with web pages. As this number is almost certainly very large, this characteristic of the method is a real benefit. In addition, if the web pages are kept in random order, the embodiments can be run on more manageable subsets of the entire set of web pages. This allows the job of finding these associations to be divided into much smaller jobs which can be run, serially or in parallel, during idle times on the server where the search engine resides. This method can produce novel associations of great width (k) at any point during its execution. Many other "association mining" methods only find longer k-tuples of associated attributes at later stages in their long execution times. Lastly, as the list of associated topics found by this algorithm grows, the pages that select the links for these new "joint topics" can be created and cached. This would reduce server loads (thus allowing more users to access the system). As this also puts bounds on the statistical relevance of the findings, this information could be used to select which new topic indices would be cached and which would be re-created as needed.

Detailed Description Text - DETX (251):

Internet and intranet search engines attempt to order the space of web pages or documents by topic. Generally, an initial (e.g. alphabetic) ordering is not at all likely to evenly divide that space. For example, the topic "California" will have a vastly greater set of pages associated with it than will "North Dakota". A simple tree-like storage of the pages by topic (with sub-topics at lower levels of the tree) will leave "California" with a very deep tree. What would be of use in this situation would be some better way to divide the search space of pages than by just single topics. In the noted example, it would be better to have the large set of California-related web pages divided into smaller sets closer to the size of the set for North Dakota. We can keep our ordering of the pages by topic if we choose to divide larger sets into smaller ones by replacing the single topic describing the set with a series of associated topic lists that encompasses the same space. Going back to our example, if "California" were only strongly associated with "Sunshine", "Wine" and "Cars" we would replace the tree node "California" with the set of nodes "California and Sunshine", "California and Wine", "California and Cars", "California and Other". This will allow faster lookup and storage of these pages because it reduces the height of this part of the tree (in this case) by one. Recursively applying the same technique at all nodes in the tree would provide a method for ensuring better balance than could have been had before. The only thing missing from this formulation of the new tree balancing function is the discovery of the associations themselves. An application of embodiments described herein to the same table discussed in the previous section extracts this information from the set of pages. The method tells us not only which topics are related but also gives an indication of the level of support for each association in the database. Once a problematically large topic has been identified, the list of associations found by the algorithm that includes this topic can be consulted to determine how to divide the topic.

Detailed Description Text - DETX (252):

The use of tree-based storage retrieval techniques is known to those in the art, and such methods include such variations as B-trees, k-D trees, tries, k-D tries, and gridfiles. Hashing schemes can also be used instead of, or in addition to, tree-based methods per se. With all such methods, there are efficiency gains to be made, in both storage (main memory and offline memory) and running time, by taking advantage of particular distributions of the data in the application domain. The embodiments described herein can, as shown above and in other ways, be used to obtain a better understanding of and exploitation of the distribution of the data.

Detailed Description Text - DETX (257):

Such questions can be addressed by the analysis of databases organized in terms of customers, transactions, demographic factors, previous marketing campaigns, and sales of particular products. For charitable organizations, the basic idea is the same, though instead of "sales" and "customers" the application is to "contributions" and "donors", for example. The principles described herein can be applied successfully to these analysis tasks, wherein one of the main current computational challenges is the discovery of associations (correlations) amongst sets of variables or attributes in very large databases. Table 8 illustrates the application to the analysis of databases on customer purchases of products. Table 9 is similar except that it illustrates the case wherein not only purchases are recorded in the data, but also information on previous marketing campaigns. Either of these schemes may be augmented by the inclusion of additional columns corresponding to demographic attributes of the customers, for example region of residence, age group, income group, gender, occupational category, and participation in community- or leisure-related activities.

Detailed Description Text - DETX (260):

Steps involved in applying the principles described herein to a sales/marketing database include: 1. Obtain sales/marketing database as described above. Where necessary, use methods known in the art to transform continuous-valued variables into discrete-state variables. 2. Present this database, in whole or part, such that each customer in the database corresponds to one or more of the M objects (rows) in the embodiment's data matrix and so that each product or mailing corresponds to an attribute (column) of the data matrix. Mailing attributes (if any) plus product attributes together comprise the N attributes (columns) in the data matrix. 3. Employ the base method above or one of the other embodiments herein on the data matrix. 4. Direct the discovered correlated k-tuples of attributes to:

Detailed Description Text - DETX (261):

A graphical viewer or printer, or A rule-generator preprocessor for rule-based system, or A report for marketing personnel, magazine/newspaper circulation directors, salespeople, managers or other users of the computer database query system, or a report-generation system, or Another computer program that performs some kind of further analysis of the data, for example, performing more in-depth statistical analysis (e.g., multiple regression) on the correlated variables, or Another computer program that performs some transformation or optimization on the database.

Detailed Description Text - DETX (267):

Use of the method on the "transpose" of the marketing database shown earlier, in order to cluster the customers is shown in Table 10. It is now the columns that correspond to a set of customers, while the rows now correspond to products purchased and demographic features. There are M' rows and N' columns, where perhaps M'=N and N'=M, for the original M and N described above. The value in table cell[j,i] is one (1) if customer i purchased product j or possesses demographic feature j and is zero (0) otherwise.

D tailed Descripti n Text - DETX (268):

Application of the Principles Described Herein to the Analysis of Medical, Epidemiological and/or Public Health

Databases

Detailed Description Text - DETX (269):

Medical scientists and practitioners have long known that many human diseases and disorders, physical and mental, are caused by complex interactions among many potential contributing factors. Such factors can include particular genetic conditions or abnormalities, exposure to biological pathogens, aspects of diet, environment (air, water, noise pollution), exposure to hazards in the home or workplace, emotional stress, substance abuse and poverty, among others. The true "causes" of a given condition often remains impossible to ascertain, though there is much folklore and anecdotal evidence offered in attempts to explain some instances. The problem of discovery and prevention of health threats is helped in recent times by the ability of researchers, insurance company representatives, epidemiologists and public health officials to compile and analyze large amounts of data on real people, healthy and sick, living and deceased. As in other applications of computers and statistical analysis to databases, one must contend in this field with a huge number of variables and the exponential complexity of their potential interactions. This kind of analysis can be improved greatly by methods that efficiently find correlations and associations amongst ten, hundreds, or thousands of variables. The principles described herein are applicable to such a situation.

Detailed Description Text - DETX (270):

Application to medical databases can also be represented in terms of the M by N data matrix we have used in other sections of this document. In one application-specific embodiment, the rows of the data matrix correspond to particular patients or subjects in a health study; and the columns correspond to factors thought to contribute to a given disease or set of diseases. Again, these factors can include socioeconomic factors, lifestyle (exercise, diet), aspects of the patient's home or workplace environment (e.g., exposure to carcinogenic chemicals), past medical treatments, and so on (See Table 11).

Detailed Description Text - DETX (273):

Steps involved in applying current invention to a medical/epidemiological/lifestyle factors database include: 1. Obtain database of medical/epidemiological/lifestyle factors as described above. Where necessary, use methods known in the art to transform continuous-valued variables into discrete-state variables. 2. Present this database, in whole or part, such that each patient/subject in the database corresponds to one or more of the M objects (rows) in the embodiment's data matrix and so that each potential disease factor corresponds to an attribute (column) of the data matrix. Additional attributes representing different diseases plus the disease factors together comprise the N attributes (columns) in the data matrix. 3. Employ the base method or other embodiments described herein on the data matrix. 4. Direct the discovered correlated k-tuples of attributes to: A graphical viewer or printer, or A rule-generator preprocessor for rule-based system, or A report for doctors, researchers, public health officials, managers or other users of the computer database query system, or a report-generation system, or Another computer program that performs some kind of further analysis of the data, for example, performing more in-depth statistical analysis (e.g., multiple regression) on the correlated variables, or Another computer program that performs some transformation or optimization on the database.

Detailed Description Text - DETX (279):

Another rather different application of the principles described herein to public health and insurance policy and practice is obtained by considering the transpose of the data matrix described above. Instead of patients as objects (rows) and potential disease factors as attributes (columns), consider what is possible when the patients correspond to columns and the factors correspond to rows. (See Table 12). Use of the current invention in this scenario produces correlated k-tuples of patients, or patient-profiles, in feature-space. This is seen to be a form of clustering of the patient data, into groups of patients or patient profiles that are roughly similar in terms of their lifestyle factors. Such clustering can be useful in designating special "low-risk" of "high-risk" types of patients or insurance applicants, to enable more optimal allocation of health services, outreach programs, insurance protection, or other resources. Once this transposition of the data is envisioned, the other steps of the preceding application to analysis of medical and other databases apply entirely analogously to the descriptions given above. (See Table 12).

Detailed Descripti n Text - DETX (280):

Use of the principles on the "transpose" of the disease factors databases shown earlier, in order to cluster the patients or policy-holders in factor-space is shown in Table 12. It is now the columns that correspond to a set of patients, medical study subjects, or potential insurance policy-holders, while the rows now correspond to potential disease factors that may include lifestyle factors, socioeconomic factors, workplace factors, and so on. There are M' rows and N' columns, where perhaps M'=N and N'=M, for the original M and N described above. The value in table cell[j,i] is one (1) if patient i possesses or has been exposed to factor j and is zero (0) otherwise.

Detailed Description Text - DETX (282):

Administrators of complex integrated systems such as computer networks and factory automation systems have been faced with the difficult diagnosis problems these system pose since their inception. Where a series of events in the system (perhaps over a protracted period of time) leads to a failure of the system as a whole, the diagnosis of the true cause of the failure can be an almost insurmountable task. For example, a network interface card on a gateway computer that fails intermittently when under high load conditions may not cause the host computer to crash but may lead to errors on other computers that use the card (by proxy) to service their network requests. Such a problem would be difficult in the extreme to track down using conventional diagnosis techniques. Tools that can present administrators with a better analysis of the conditions on the system as a whole that lead to the failure would speed the diagnosis and correction of the underlying problem.

Detailed Description Text - DETX (283):

We need to define the database upon which the principles described herein will be applied.

Detailed Description Text - DETX (284):

The database as a whole can be thought of as a state record of a series of components over time. The columns of this database, when viewed in the data matrix format used throughout this document, represent the series of components; the rows represent discrete points in time. The values in the table are intended to be an encoding of each component's state (on, off, idle, error, and so on) at the time in question. Such logging procedures are well known to those skilled in the art.

Detailed Description Text - DETX (286):

Steps involved in applying the method of the current invention to analysis of a system operations database include: 1. Create a database of system components and their states as described above. The choice of state sets for the components in the system will be driven by behaviors of interest to the administrators of the system as well as by the components themselves. 2. Present this database, in whole or part, as a data matrix such that each column in the data matrix corresponds to a component in the system and each row in the data matrix corresponds to a point in time in the series. 3. Employ the base method above or one of the other embodiments described herein on the data matrix. 4. Direct the discovered correlated k-tuples of attributes to: A graphical viewer or printer, or A rule-generator preprocessor for rule-based system, or A report for the administrators of the system, or a report-generation system, or Another computer program that performs some kind of further analysis of the data, for example, performing more in-depth analysis on the correlated variables, or

Detailed Description Text - DETX (287):

The output in this application, can be used to indicate the events in the system that are typically seen to co-occur with a given failure. Given the formulation of the database, we need not restrict ourselves to the states of the components in the system at the time of the failure--we can expand our examination of the failure conditions to any range of points in time for which the database has records. This allows the method to help illuminate subtle causal relationships between components that ultimately lead to failure. In the simplest case, the output can be used to eliminate some components in the system from scrutiny if it is seen that they are not correlated with the failure.

Detailed Description Text - DETX (289):

Complex systems define a large family of somewhat similar applications. For the purpose of this discussion, complex systems are defined as systems for which three are no direct detailed modeling approaches because these systems comprise a huge number of interacting individual components or parts. Examples would include (but would not be limited to) economics, individual human behavior, productivity in groups of employees, weather patterns, crime in a nation, etc. In each of these cases, there are no known methods to model the system exactly so variables or sets of variables are used to measure the state of these systems (examples in the case of economics would be the interest rate, stock market values and inflation rates). For the purposes of this description, the events in these complex systems take the form: pre-condition, action and post-condition. These interactions represent the state of the system before the actions were taken, the actions themselves and the resulting state of the system at some point after the implementation of the actions. Put another way, the set of previous perturbations of the system and their outcomes are used as a history of the system from which to derive information about the system's characteristics.

Detailed Description Text - DETX (290):

The kinds of databases of complex systems that can effectively utilize the principles described herein must meet certain restrictions. There must be some set of variables (either in common usage or derivable from knowledge in the domain) used to measure the state of the given system. These variables are used in the pre and post condition parts of each database entry. Additionally, there must be some general set of actions that may be applied to the system that encompass methods by which it is known the system may be perturbed. Returning to the economics example, the action set would include all things under the heading of "fiscal policy".

Detailed Description Text - DETX (291):

Formally, the database must include attributes representing zero or more pre-condition variables zero or more action variables, and zero or more post-conditions variables. Leaving aside the trivial case wherein the database contains zero pre and post condition variables and zero action variables, there are eight cases to consider. They will be presented exhaustively below with examples where appropriate. Note that in each case, there are two interpretations of relevance. For example, consider the case where we have pre-condition variables and action variables but no post-conditions. The correlations can be derived in two ways: the database itself could have had no post-condition variables in it (and the returned set of correlations is culled to remove any correlations that involved only variables of one type) or it can be that just the set of correlations themselves contain no post-condition variables even though the database does in fact contain them. For the purposes of the discussion, we assume the former is the case--we can always cull the results of the method on a database that has more types of variables to leave a set of correlations which do not have some types of variables.

Detailed Description Text - DETX (292):

If the database contains only variables of one type (i.e. only action variables or pre or post condition variables) then the correlations derived from it can be interpreted in one of two ways. If the variables are pre or post condition variables, then the results indicate situational archetypes--that is, sets of attribute values (or, equivalently, states of variables) that tend to be seen together. An example from the domain of weather patterns would be rain and low barometric pressure. If only action variables are present in the database then correlations found between them indicate sets of decisions that tend to be made together. In a military domain, we might discover that flanking maneuvers and offensives tended to be seen co-occurring. As these types of databases are very similar to others described elsewhere in this document (as would be the applications of the method in these cases), this section will not explicitly address them.

Detailed Description Text - DETX (293):

The cases where the database contains variables of only two of the three types are three in number.

Detailed Description Text - DETX (294):

Correlations found in a database that contains only pre-condition and action variables describe the relationship between situations in the domain and the selection of actions. An example is football play-calling (note that this also involves a complex system that can not be modeled in any direct detailed way--the play-caller). Here the correlations indicate the tendencies of the action-taking entity, e.g., a coach or quarterback.

Detailed Description Text - DETX (295):

If the database contains only action and post-condition variables, then the correlations found elucidate the effectiveness of sets of actions regardless of pre-conditions. Going back again to the football example, correlations of this type would illuminate the ability of the team in question to perform certain actions (e.g., if "third and long yardage to first down" tended to result in a poor post-condition set, like fourth down, then we would know that the team tended to be ineffective in this situation). Another important example is drug interaction. In this case, the actions are the drugs given and the post-conditions are the side-effects reported for some patient.

Detailed Description Text - DETX (296):

While the utility of the case where the database contains only pre and post condition variables may be unclear on first examination, it may well be that this is one of the most useful cases. Here we are either interested in things that tend to happen after a situation in the given domain regardless of actions taken by the decision-maker or we are in a domain where there are no actions that can be taken (or none that effect the system itself). An example of the former would be the fact that the pre-condition "third and long" in football tends to be followed by the post-condition "fourth and long". In fact, it may be the latter case that is the most interesting. Consider that case of weather patterns. If we focus on the post-condition "tornadoes" (that is, we cull the resulting correlation set so that it includes only those correlations that involve the appearance of "tornadoes" in the post-condition), then what these correlations tell us are precursor signs that tornadoes are immanent.

Detailed Description Text - DETX (297):

The last case is the most general: the database contains all three types of variables. Note that a database of this form is capable of having correlations of attributes of all the preceding types. Example domains have already been given (economies, crime in a population, etc.) Here the correlations can be thought of as rating actions sets (given some set of pre-conditions) based on the quality of the post-conditions.

Detailed Description Text - DETX (298):

The last consideration is the types of data that the database entries contain. Binary valued attributes, as noted throughout this document, can readily be accepted by this method. Other value types must be of limited range of discrete values. Where this is not the case (i.e. real-valued or integer-valued attributes), some transformation must be performed on the values in question to reduce their range of values to a more manageable number. Various clustering methods are among the preferred methods for this, and are well-known to those skilled in the art.

Detailed Description Text - DETX (301):

Description of the Application of the Principles Described Herein to Databases with Pre-condition Variables and Action Variables:

Detailed Description Text - DETX (302):

Given the above-noted restrictions on the form of the database, it is clear that the input requirements for the application of the embodiments described elsewhere herein are met. In the convenient data matrix representation cited elsewhere in this document, the M rows in this context are the total selected set of pre-conditions and actions taken. If the entity that applies the actions can sensibly be personified then these rows can represent a history of the decisions made by the entity and the states of the system at the time they were made. The N columns comprise the set of state variables that define the state of the system and the set of all applicable action variables that describe the ways in which the system can be perturbed (see Table 14).

Detailed Description Text - DETX (305):

Previously noted examples are the case of football play-calling by coaches and military decision made by generals. In general, preferred implementations of this invention will use the method of the current invention on databases of this form in order to extract information about the action-taking entity. The correlated state variables and actions describe the tendencies of this entity. As noted above, these may be further analyzed using case-based reasoning tools to give a better picture of the entity's likely decisions given a state of the system.

Detailed Descripti n Text - DETX (306):

Another use of the invention on databases of this type is in discovering fraud indicators in tax collection. Here we let the pre-conditions be a set of attributes intended to capture the salient details of a tax return (such things as total income, total tax owing as reported by the individual or business, tax exemptions claimed, etc.) and choose the action variables to define a set of possible tax evasion methods. The correlations found by the invention then indicate associations between types of tax returns and types of tax evasion. As coincidence detection bounds the returned correlations statistically, we not only find indicators of evasion but also the reliability of these findings. Given that tax collection agencies can not afford to investigate all tax returns sent to them, this method allows them to find a well-chosen subset of these returns that is most likely to result in findings of fraud (and greater monetary returns for the government).

Detailed Description Text - DETX (308):

Steps involved in applying thprinciples described herein to a database containing pre-condition and action variables include: 1. Create the database of system states and actions taken by the action taking entity as described above. Where necessary, use methods known in the art to transform continuous-valued attributes into discrete-state attributes. 2. Present this database, in whole or part, such that each states/action set corresponds to one of the M objects (rows) in a data matrix and so that each state type aspect and action type corresponds to an attribute (column) of the data matrix. 3. Employ the base method or other embodiment described herein on the data matrix. 4. Direct the discovered correlated k-tuples of attributes to: A graphical viewer or printer, or A report for decision-makers, or a report-generation system, or Another computer program that will use the correlations found as a basis for making decisions (for example, a case-based reasoning package), or Another computer program that performs some transformation or optimization on the database.

Detailed Description Text - DETX (310):

Description of the Principles Described Herein as Applied to Databases with Pre-condition Variables and Post-condition Variables:

Detailed Description Text - DETX (311):

Here, too, the above-noted restrictions on the form of the database force compliance with the input requirements of the embodiments described elsewhere herein. The M rows in this context are the instances or combinations of pre-conditions and post-conditions (viewed together, one can think of these rows as being the system's transitions between states). The N columns are comprised of the set of state variables that define the state of the system before and after the transition (see Table 15).

Detailed Description Text - DETX (312):

The value in cell[i,j] of Table 15 is an encoding of the measure of state variable j either before or after the transition.

Detailed Description Text - DETX (314):

Equally important is the selection of time quanta that define the granularity of the transitions. This too is left to those skilled in the art to decide based on their own expertise and the kinds of information they wish to extract. It is assumed that some minimum granularity is imposed by either the complexity of gathering such data or by the limits of the usefulness of such data. Given this, one can then pick any multiple of this minimum granularity to be the time between pre and post conditions. At the very least, this distance in time should be long enough for the system to have changed its state.

Detailed Descripti n Text - DETX (316):

In the domain of economics and fiscal policy, we propose a database of sets of states where the states are a set of economic indicators (inflation and interest rates, housing starts, GDP and so on). Each row in the database should contain two such states (the pre and post condition of the system) separates by a fixed amount of time. The correlations found in by the method of the current invention then give insight into cycles in the economy.

Detailed Descripti n Text - DETX (317):

For stock market prediction, we propose a set of stocks (presumably large) which are thought to have influence over one another. Again, a fixed period of time is selected for transitions. The rows of this database then tell the transition of these stocks over the chosen period of time. The output of the invention then indicates which sets of stocks "move" in a correlated manner over that period of time.

Detailed Description Text - DETX (320):

In general, preferred embodiments of this invention will use the method of the current invention on databases of this form in order to extract information about how the current state of the system acts as a predictor for a future state. Given probabilistically bounded data correlations between states of the system, effective predictions can be made about the system's behavior.

Detailed Description Text - DETX (321):

Steps involved in applying current invention to a database containing pre-condition and action variables include: 1. Create the database of transitions between system states, wherein a system state is represented by a value of a state variable, over the chosen time quantum as described above. Where necessary, use methods known in the art to transform any continuous-valued state variables into discrete-state variables. 2. Present this database, in whole or part, such that each state to state transition set corresponds to one of the M objects (rows) in the embodiment's data matrix and so that each state variable corresponds to an attribute (column) of the data matrix. 3. Employ the base method or other embodiment described herein on the data matrix. 4. Direct the discovered correlated k-tuples of attributes to: A graphical viewer or printer, or A report for decision-makers, or a report-generation system, or Another computer program that will use the correlations found as a basis for making decisions (for example, a case-based reasoning package), or Another computer program that performs some transformation or optimization on the database.

Detailed Description Text - DETX (322):

Description of the Application of the Principles Described Herein to Databases with Action Variables and Post-condition Variables:

Detailed Description Text - DETX (323):

Here, too, the above-noted restrictions on the form of the database force compliance with the input requirements of the embodiments described elsewhere herein. The M rows in this context are the total selected set of actions and post-conditions. The N columns are comprised of the set of state variables that define the state of the system before and after the transition (See Table 16).

Detailed Description Text - DETX (324):

The rows of Table 16 correspond to observed instances of, or hypothetical combinations of, actions applied to the system and their resulting system states. The columns correspond to either possible actions that can be applied to the system or are individual state representation variables. If column p corresponds to one of the action types in the database, the value in table cell[i,p] of Table 16 is an encoding of the action taken. If column j is a column used to indicate some aspect of a state of the system, then the value in the table cell[i,j] is an encoding of the measure of that aspect.

Detailed Description Text - DETX (325):

As noted in previous examples, decisions that must be made prior to the application of the method of the current invention to databases of this type include the choice of state variables used to store the state of the system at a given point in time and the choice of time quantum used to temporally separate the actions from the post-conditions. These choices are left to those skilled in the domain of application. The time quantum chosen must, in the most trivial case, be long enough for the actions to have had some effect on the state of the system.

Detailed Descripti n Text - DETX (329):

Steps involved in applying current invention to a database containing action and post-condition variables include: 1. Create the database of transitions between system states and actions over the chosen time quantum as described above, wherein a system state is represented by a value of a state variable and an action is represented by a value of an action type. Where necessary, use methods known in the art to transform continuous-valued state variables and action types into discrete state variables and action types. 2. Present this database, in whole or part, to an embodiment of the current invention such that each action set/state set pair corresponds to one of the M objects (rows) in the embodiment's data matrix and so that each state variable or action type corresponds to an attribute (column) of the data matrix. 3. Employ the base method or other embodiment described herein on the data matrix. 4. Direct the discovered correlated k-tuples of attributes to: A graphical viewer or printer, or A report for decision-makers, or a report-generation system, or Another computer program that will use the correlations found as a basis for making decisions (for example, a case-based reasoning package), or Another computer program that performs some transformation or optimization on the database.

Detailed Description Text - DETX (330):

Description of the Application of the Principles Described Herein to Databases with Pre-condition Variables, Action Variables and Post-condition Variables:

Detailed Description Text - DETX (331):

Here, too, the above-noted restrictions on the form of the database force compliance with the input requirements of the embodiments described elsewhere herein. The M rows in this application are the total selected set of pre-conditions, actions and post-conditions. The N columns are comprised of the set of state variables that define the state of the system before and after the transition as well as the encoded actions types (see Table 17).

Detailed Description Text - DETX (332):

The rows of Table 17 correspond to instances or combinations of pre-condition, actions taken and the resulting post-conditions. The columns correspond to types of actions possible in the domain as well as aspects of interest to any given situation in the domain (for both pre and post condition columns). If column p corresponds to one of the action types in the database, the value in cell[i,p] of Table 17 is an encoding of the action taken. If column p is a column used to specify some aspect of either the pre-condition or the post-condition, then the value in table cell [i,j] is an encoding of the measure of that aspect.

Detailed Description Text - DETX (333):

As noted in previous examples, decisions that must be made prior to the application of the method of the current invention to databases of this type include the choice of state variables used to store the state of the system at a given point in time and the choice of time quantum used to temporally separate the actions from the post-conditions. In this case, it should be noted that it is not necessary for the pre and post-conditions to be equivalent (with respect to the choices of variables). These choices are left to those skilled in the domain of application. The time quantum chosen must, for example, be long enough for the actions to have had some effect on the state of the system.

Detailed Description Text - DETX (335):

Given some set of variables to define the state of an economy (interest rates, inflation, GNP and so on) and a set of actions taken as part of the governing body's economic policy (issuing and buying back government bonds, etc.), we create a database of economic events of the form: existing economic state, fiscal policy measures taken and economic state following the policy decisions. The correlations found by the method of the current invention give a measure to the effectiveness of economic policy decisions, given a state of the economy. Such knowledge would be beneficial in deciding economic policy as it would show historical support (or the lack thereof) for a given set of decisions.

Detailed Description Text - DETX (336):

In a similar vein, the use of the current invention to aid in setting anti-crime policy starts with the creation of a database of previous states of the community's crime, policy measures taken and the resulting state of crime in the community. The state variables could include things like the rates for differing types of crimes (breaking and entering, auto theft, etc.), differing characteristics of crimes (i.e. whether or not handguns were used etc.) and so on. The action variables in this case could include such things as minimum sentencing guidelines for various crimes, "three-strike" laws, the adoption of the death penalty, as well as education and mental health funding. On such a database, the invention would find correlations involving existing crime states, policy decisions and the outcomes of those decisions. It is proposed that these correlations could prove an invaluable aid to those charged with making such decisions.

Detailed Description Text - DETX (337):

The concept of the "decision-maker" needs careful consideration in the domain of military strategy. It may well be the case that there is not enough of a "track record" to fill a database with enough of a history of any one general's decision making. In such a case, preferred implementations can extend the concept of the decision-maker to include all similar decision-makers. As an example, consider a single general commanding a tank division. If the general were recently promoted, one would be wise to consider all the history of all such generals of the same allegiance. To increase further the granularity of the use of the method, the database could be filled with the decisions made by all infantry lieutenants rather than with those of any one lieutenant. Correlations found would be indicative of the tendencies of that class of generals given some measure of the battlefield conditions faced when they made their decisions. Equally, one would be in a position to determine which battlefield situations they handled poorly because one has access to the outcomes of the decision sets. Such knowledge could prove vital to selecting an opposing strategy.

Detailed Description Text - DETX (338):

Steps involved in an application of the principles described herein to a database containing pre-condition, action and post-condition variables include: 1. Create the database of states and actions covering the chosen time quantum as described above. Where necessary, use methods known in the art to transform continuous-valued state variables and action types into discrete state variable and action types. 2. Present this database, in whole or part, such that each state/action/state triple corresponds to one of M objects (rows) in a data matrix and so that each state variable or action type corresponds to an attribute (column) of the data matrix. 3. Employ the base method or other embodiment described herein on the data matrix. 4. Direct the discovered correlated k-tuples of attributes to: A graphical viewer or printer, or A report for decision-makers, or a report-generation system, or Another computer program that will use the correlations found as a basis for making decisions (for example, a case-based reasoning package), or Another computer program that performs some transformation or optimization on the database.

Detailed Description Paragraph Equation - DEEQ (1):

P(Observed(a.sub.i1 @c.sub.i1, a.sub.i2 @c.sub.i2, . . . , a.sub.ik @c.sub.ik).vertline.Independent(c.sub.i1, c.sub.i2, . . . , c.sub.ik), Model)<.theta.,

Claims Text - CLTX (1):

1. A coincidence detection method for use with a data set of objects, said objects having a number of attributes, the method comprising the steps of: sampling various subsets of the data set for a plurality of iterations, each iteration the sampled subset of the data set having for each object the same subset of attributes; detecting, and recording counts of, coincidences in each sampled subset of the data set, a coincidence being the co-occurrence of a plurality of attribute values in one or more objects in a sampled subset of the data set, where the plurality of attribute values is the same for each occurrence, the detecting and recording counts of coincidences in each

sampled subset of the data set being performed before, at the same time or after sampling, detecting and recording counts of coincidences in other subsets; determining an expected count for each coincidence of interest that has been detected in the previous step; comparing, for each coincidence of interest, the observed count of coincidences versus the expected count of coincidences, and from this comparison determining a measure of correlation for the plurality of attributes for the coincidence; and reporting a set of k-tuples of correlated attributes, where a k-tuple of correlated attributes is a plurality of attributes for which the measure of correlation is above a respective pre-determined threshold.

Claims Text - CLTX (3):

3. The coincidence detection method of claim 1, wherein the counts are recorded by storing a running total of the count of each coincidence over all of the sampled subsets.

Claims Text - CLTX (8):

7. The method of claim 1, further comprising the steps of first creating a database of transitions between system states, wherein a system state is represented by a value of a state variable, over a chosen time quantum, and presenting the database, in whole or part, as a data set such that each state to state transition set corresponds to one of the objects and so that each state variable corresponds to an attribute.

Claims Text - CLTX (9):

8. The method of claim 1, further comprising the steps of first creating a database of states and actions covering a chosen time quantum and presenting the database in whole or part, as a data set such that each state/action/state triple corresponds to one of the objects and so that each state variable or action type corresponds to an attribute.

Claims Text - CLTX (19):

18. A coincidence detection method for use with a data set of objects, each of the objects having at least one attribute, the method comprising the steps of: (1) sampling various subsets of the data set for a plurality of iterations, each iteration the sampled subset of the data set having for each object the same subset of attributes; (2) detecting attribute coincidences based on results of sampling of the data set; (3) recording attribute coincidences; and (4) comparing at least one recorded attribute coincidence count to at least one expected attribute coincidence count, wherein the expected attribute coincidence count is determined for a coincidence that has been detected in the preceding steps.

Claims Text - CLTX (24):

23. The method of claim 18, the method further comprising the step of separating the data set into subsets for sampling.

Claims Text - CLTX (25):

24. The method of claim 18, wherein more than one subset of objects is sampled and the size of the subsets of objects sampled is a constant.

Claims Text - CLTX (27):

26. The method of claim 18, wherein step (3) comprises the step of storing a running total of counts of each attribute coincidence detected over all the subsets sampled.

Other Reference Publication - OREF (7):

A.F.W. Coulson et al., "Protein and nucleic acid sequence database searching: a suitable case for parallel processing," The Computer Journal, vol. 30, No. 5, Oct. 1987, Cambridge, Great Britain, pp. 420-424.

Other Reference Publication - OREF (9):

R. Guigoet al., "Inferring Correlation between Database Queries: Analysis of Protein Sequence Patterns," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, No. 10, Oct. 1993, New York, New York, USA, pp. 1030-1041.

Other Reference Publication - OREF (14):

A. Krogh et al., "Hidden Markov Models in Computational Biology: Applications to Protein Modeling," J. Mol. Biol., vol. 235, 1994, pp. 1501-1531.

Other Reference Publication - OREF (16):

C. de Marcken, "Unsupervised Language Acquisition," Ph.D. Thesis, M.I.T. (Sep. 1996) (title page, abstract, and pp. 82-93).

Other Reference Publication - OREF (20):

Steeg, E. et al., "Application of a Noval and Fast Information-Theoretic Method to the Discovery of Higher-Order Correlations in Protein Databases" in Proceedings of the 1998 Pacific Symposium on Biocomuting. Ed. Altman et al., World Scienctific Publishing Co., New Jersey, pp. 573-584 (Jan. 4-9, 1998).